Amendments to the Claims

This listing of claims will replace all prior versions, and listings of claims in the application.

1. (Currently Amended) A computer-based method for representing latent semantic content of a plurality of documents, each document containing a plurality of terms, the method comprising:

deriving at least one n-tuple term from the plurality of terms;

forming a two-dimensional matrix,

each matrix column c corresponding to a document,

each matrix row r corresponding to a term occurring in at least one document corresponding to a matrix column,

each matrix element (r, c) related to a number of occurrences of the term[[;]] corresponding to the row r in the document corresponding to column c,

at least one matrix element related to the number of occurrences of at least one n-tuple term occurring in the at least one document, and

performing singular value decomposition and dimensionality reduction on the matrix to form a latent semantic indexed vector space and storing the latent semantic indexed vector space in an electronic form accessible to a user.

2. (Currently amended) The computer-based method as recited in claim 1 further comprising:

identifying an occurrence threshold;

wherein n-tuples that appear less times in the document collection than the occurrence threshold are not included as elements of the matrix.

3. (Previously presented) The computer-based method as recited in claim 2 wherein the occurrence threshold is two.

4. (Previously presented) The computer-based method as recited in claim 1 wherein deriving at least one n-tuple term further comprises:

creating the at least one n-tuple term from n consecutive verbatim terms.

5. (Previously presented) A computer-based method for determining conceptual similarity between a subject document and at least one of a plurality of reference documents, each reference document containing a plurality of terms, the method comprising:

deriving at least one n-tuple term from the plurality of terms;

forming a plurality of two-dimensional matrices wherein, for each matrix:

each matrix column c corresponds to a document, wherein one column corresponds to the subject document and the remaining columns correspond to the reference documents;

each matrix row r corresponds to a term occurring in at least one of the subject document or the reference documents,

each matrix element (r, c) represents a number of occurrences of the term corresponding to r in the document corresponding to c;

performing singular value decomposition and dimensionality reduction on the plurality of formed matrices, to form a plurality of latent semantic indexed vector spaces,

the plurality of latent semantic indexed vector spaces including at least one space formed from a matrix including at least one element corresponding to the number of occurrences of at least one n-tuple term in at least one document,

determining at least one composite similarity measure between the subject document and the at least one reference document as a function of a weighted similarity measure of the subject document to the at least one reference document in each of the plurality of indexed vector spaces and storing the at least one composite similarity measure in an electronic form accessible to a user.

6. (Currently amended) The method as recited in claim 5 wherein determining the at least one composite similarity measure comprises weighing similarity measures from vector spaces comprising greater numbers of n-tuples greater than similarity measures from vector spaces comprising lesser numbers of n-tuples.

7-13 (canceled).

14. (Currently amended) A computer-based method for characterizing results of a query comprising:

automatically identifying n-tuples included in a collection of documents based on an analysis of the collection of documents, wherein each document in the collection of documents contains a plurality of terms;

forming a latent semantic indexed vector space based on (i) the documents in the collection of documents, (ii) the plurality of terms, and (iii) the automatically identified n-tuples;

querying the latent semantic indexed vector space with a query having at least one term;

ranking results of the querying step as a function of at least a frequency of occurrence of the at least one term, thereby generating a characterization of the results; and

storing the characterization in an electronic form accessible to a user.

15. (Original) The method as recited in claim 14 wherein at least one term used in ranking is a query term.

16. (Original) The method as recited in claim 15 wherein the at least one query term used in ranking is a generalized entity.

17. (Original) The method as recited in claim 14 wherein the at least one term used in ranking is a generalized entity.

18-21 (canceled).

22. (Currently amended) A computer-based method for representing latent semantic content of a plurality of documents, each document containing a plurality of verbatim terms, the method comprising:

deriving at least one expansion phrase from the verbatim terms,

each expansion phrase comprising terms;

replacing at least one occurrence of a verbatim term having an expansion phrase with the expansion phrase corresponding to that verbatim term;

forming a two-dimensional matrix,

each matrix column c corresponding to a document;

each matrix row r corresponding to a term[[()]];

each matrix element (r, c) representing a number of occurrences of the term corresponding to r in the document corresponding to c;

at least one matrix element corresponding to the number of occurrences of [[one]] at least one term occurring in the at least one expansion phrase, and

performing singular value decomposition and dimensionality reduction on the matrix to form a latent semantic indexed vector space and storing the latent semantic indexed vector space in an electronic form accessible to a user.

23. (Currently amended) A computer-based_method for representing [[the]] latent semantic content of a plurality of documents, each document containing a plurality of terms, the method comprising:

identifying at least one idiom among the documents,

each idiom containing at least one idiom term;

forming a two-dimensional matrix,

each matrix column corresponding to a document;

each matrix row corresponding to a term occurring in at least one document represented by a row;

each matrix element representing a number of occurrences of the term corresponding to the element's row in the document corresponding to element's column;

at least one occurrence of at least one idiom term being excluded from the number of occurrences corresponding to that term in the matrix,

performing singular value decomposition and dimensionality reduction on the matrix to form a reduced matrix and storing the reduced matrix in an electronic form accessible to a user.

24. (Currently amended) A computer-based method for representing [[the]] latent semantic content of a plurality of documents, each document containing a plurality of terms, the method comprising:

identifying at least one idiom among the documents,

each idiom containing at least one idiom term;

replacing at least one identified idiom with a corresponding idiom elaboration, each elaboration comprising at least one elaboration term,

forming a two-dimensional matrix,

each matrix column corresponding to a document;

each matrix row corresponding to a term;

each matrix element representing a number of occurrences of the term corresponding to the element's row in the document corresponding to element's column,

at least one matrix element corresponding to the number of occurrences of an elaboration term in a document corresponding to a matrix column;

performing singular value decomposition and dimensionality reduction on the matrix to form a reduced matrix and storing the reduced matrix in an electronic form accessible to a user.